Intra- and inter-rater reliability of the modified ENT assessment score (ENTAS 2) in granulomatosis with polyangiitis: a prospective randomised trial

L. Decker[1], L. Türp[1], C. Borzikowsky[2], M. Laudien[1]

[1]Department of Otorhinolaryngology, Head and Neck Surgery, University of Kiel; [2]Institute of Medical Informatics and Statistics, University Medical Center SH, Campus Kiel, Germany.

*Lars Decker, MD**
*Lisa Türp, MD**
Christoph Borzikowsky, MD, PhD
Martin Laudien, PhD

**These authors contributed equally to this work.*

Please address correspondence to:
Dr M. Laudien,
Department of Otorhinolaryngology, Head and Neck Surgery,
Kiel University,
Arnold-Heller-Straße 3,
24105 Kiel, Germany,
E-mail: laudien@hno.uni-kiel.de

Received on December 28, 2016; accepted in revised form on March 27, 2017.

Clin Exp Rheumatol 2017; 35 (Suppl. 103): S59-S66.

Key words: vasculitis, granulomatosis with polyangiitis, Wegener's granulomatosis, ears nose and throat (ENT), disease activity, diagnostic imaging, endoscopy

Competing interests: none declared.

ABSTRACT

Objective. *Ears nose and throat (ENT) involvement is found on a substantial proportion of patients with granulomatosis with polyangiitis (GPA). Structured, reliable ENT assessment is essential in the management of GPA patients. It is the aim of this study to determine the repeatability (intra-rater reliability) and reproducibility (inter-rater reliability) of the ENT Assessment Score (ENTAS 2).*

Methods. *The ENTAS 2 built the fundament of the prospective randomised trial. Anamnestic, video endoscopic and diagnostic data of 47 patients were used. A single assessor reference was created. GPA/ENT activity and damage were evaluated by three physicians at two time points (T1/T2). GPA/ENT activity was evaluated in dichotomy (yes/no) and grading (none/mild/moderate/high) and GPA/ ENT damage in dichotomy.*

Results. *ENTAS 2 activity evaluations intra-rater reliability was 80.7% (κ=0.56) in dichotomy and 72.8% (κ=0.41) in grading. ENTAS 2 damage evaluations showed 87.8% (κ=0.74) intra-rater reliability. ENTAS 2 activity inter-rater reliability at T1 was 62.2% (κ=0.43) in dichotomy and 51.1% (κ=0.29) in grading, at T2 it was 68.2% (κ=0.48) in dichotomy and 55.32% (κ=0.33) in grading. Inter-rater reliability of ENTAS 2 damage evaluation was 84.4% (κ=0.79) at T1 and 72.5% (κ=0.64) at T2.*

Conclusion. *ENTAS 2 intra-rater reliability was high in dichotomous and graded GPA/ENT activity and damage evaluations. Inter-rater reliability was high in dichotomous activity and damage evaluations, but low in graded activity evaluations. The data demonstrate that the ENTAS 2 is a reliable score-system considering GPA/ENT activity and damage evaluations.*

Introduction

Granulomatosis with polyangiitis (GPA, formerly known as Wegener's granulomatosis) is a systemic disease characterised by a small to medium vessel vasculitis and a necrotising granulomatous inflammation. The GPA is often associated with the presence of proteinase 3-specific cytoplasmic anti-neutrophil cytoplasmic antibodies (PR3-cANCA) and predominantly affects the upper respiratory tract, the lung and the kidney (1). Recently, significant improvement has been achieved in understanding the immunologic and genetic background of the disease, as well as in the therapeutic options (2). The current recommendation by the European League Against Rheumatism (EULAR) and the European Renal Association-European Dialysis and Transplant Association (ERA-EDTA) suggests a patient individual therapy combination of immunosupressants and/or biologicals (3). To state therapy suggestions, disease activity has to be defined. High disease activity results in an escalation of the therapy, whether the abstinence of activity can be seen as remission and lead to a de-escalation of the therapy (4, 5). Score-systems to estimate activity have been created (6-8). The most common system is the Birmingham Vasculitis Activity Score (BVAS) (5). The modified BVAS version 3 (BVAS v.3) is a tool that showed convergent validity towards rheumatologist treatment decisions (4). Periods of active disease can lead to multisystemic damages. These damages are commonly described and evaluated by the Vasculitis Damage Index (VDI) (9).

Otorhinolaryngological assessment is essential in the monitoring of GPA patients. ENT manifestations are the first symptoms in >60% of the patients with GPA (10). Often, damage in the ENT region caused by a high disease activity

requires special treatment. Saddle noses can lead to nasal reconstructive surgery (11, 12). Subglottic stenoses require special treatment such as dilatation with or without topical corticosteroid or mitomycin C, laser surgery or cryotherapy to avoid tracheostomy (13-17).

There have been trials to create GPA scoring systems specialised in the head and neck region (18, 19). The ENT activity score (ENTAS first version) showed high intra- and inter-rater reliability considering dichotomous (yes/no) evaluations of endonasal GPA activity (20). However, the repeatability (intra-rater reliability) and reproducibility (inter-rater reliability) of a GPA activity and damage score evaluating a full assessment of the ENT region have not been validated yet. ENT endoscopy is recommended to detect GPA mucosa affections (1). However, the impact of endoscopy based ENT activity and damage evaluations within a GPA score-system has not been validated yet.

It is the aim of the present study to determine the intra- and interrater reliability of the slightly modified ENT assessment score (ENTAS 2). Determining the multi-method agreement between video endoscopy based and ENTAS 2 based activity and damage evaluations is another main aspect of this study.

Patients and methods

Data collection

The ENTAS first version was modified by the use of the TRIAD ENT Database (20, 21). The modified ENTAS 2 is shown in Figure 1. Data were collected on the basis of the ENTAS 2: Rhinopharyngo-laryngoscopies, otoscopies and endoscopic inspections of the oral cavity were recorded as digital videos. Anamnestic, video and diagnostic data (audiometry, Rinne/Weber Test, tympanometry, rhinomanometry and sniffing sticks test) of initially 135 patients with GPA were prospectively collected from 10/2013 until 05/2014. Examples of endoscopic findings are shown in Figure 2. Subsequently, the videos were analysed in quality (integrity of all ENT sections, sharpness, brightness, stability of camera work, period of possible mucosa evaluation). Patients with insufficient image quality were excluded.

ENT ASSESSMENT SCORE (ENTAS 2)

Name:
Date of Birth:
Diagnosis:

1.1 Activity symptoms
Subjective complaints (new appearance or worsening; not explained by other conditions)

Pain

	Yes	No
Nose		
Paranasal sinus		
Headache		
Oral cavity		
Odynophagia		
Larynx/throat		
Ear		

Functional impairment

	Yes	No
Nasal obstruction		
Olfactory loss		
Sudden hearing loss (<72h)		
Chronic hearing loss		
Dizziness		
Hoarseness/ *Voice change*		
Dyspnea		
Dysphagia		
Tinnitus		
Vertigo		

Secretion

	Yes	No
Otorrhea		
Epistaxis		
Rhinorrhea		

1.2 Physical findings

Localisation	Clinical findings
Skin	( ) absent ( ) erythema ( ) ulcers
Nose	( ) absent ( ) friable mucosa ( )oedema ( ) ulcers ( )granulation ( )crusts
Epi-/ Oropharynx	( ) absent ( ) oedema ( ) ulcers ( )granulation ( ) bloody mucosa
Hypopharynx (h), Larynx (l), Trachea (t)	( ) absent ( ) friable mucosa ( )oedema ( ) ulcers ( )granulation ( ) stenosis
Subglottic Stenosis	*( ) absent ( ) <50% ( ) 51-70% ( ) 71-99% ( ) 100% ( ) Circumfer. ( ) Spiral ( ) Web*
Ear	( ) absent ( ) friable mucosa ( )oedema ( )ulcers ( )granulation ( ) stenosis ( ) secretion *( ) tympanic perforation ( ) glue ear ( ) inflammation ( ) retracted drum ( ) grommet in situ ( ) tympanosclerosis*
Oral cavity	( ) absent ( ) oedema ( ) ulcers ( )granulation ( ) bloody mucosa
Objective stridor	( ) absent ( ) present

1.3 Other findings
(Cranial nerves, lymph nodes, glands, neck mass etc)

1.4 Diagnostic tests

	Diagnostic findings
Audiometry	( ) normal ( ) CHL >10dB ( ) SHL* 20-40dB ( ) SHL*40-70dB ( ) SHL*>70dB
Impedance	( ) normal ( ) effusion ( ) perforation
Sniffin' sticks	/12
SDI	
Trigeminus damage nose	*( ) yes ( ) no*
Rhinomanometry	nasal obstruction: ( ) absent ( ) mild ( ) moderate ( ) severe

 # new or worsened and not explained by other conditions (SHL: sensorineural hearing loss; CHL: conductive hearing loss) *at 1, 2 or 3 kHz

	None	Mild	Moderate	High
Activity grading				
Treatment proposal (+ acceleration, - tapering, = unchanged)				

Procedure/Suggestions:

2.0 Damage (considered to be caused by Vasculitis)

Septal perforation		Saddle nose		Synechia (endonasal)		Laryngeal/tracheal stenosis		Cranial nerve palsy	
Ocular symptoms		SHL		Cervical lymph nodes		Tympanic perforation		Olfactory distortion	

Date: Examiner: Supervisor:

Fig. 1. ENT Assessment Score (ENTAS 2).

Finally, the data of 47 GPA patients were used in the study. Patients gave written informed consent and the study was approved by the local ethics committee. The data were anonymised and randomised.

Cohort parameters

GPA was defined following the EULAR recommendations (5) and classified by the 2012 Revised International Chapel Hill Consensus Conference nomenclature and the American College of Rheumatology criteria (22, 23). 47 patients with GPA (30 women; 17 men) were included in the study. Mean age of the patients was M=56 years (range 20-85; SD=15). GPA disease stages (localised; early systemic; generalised; severe; refractory) were defined

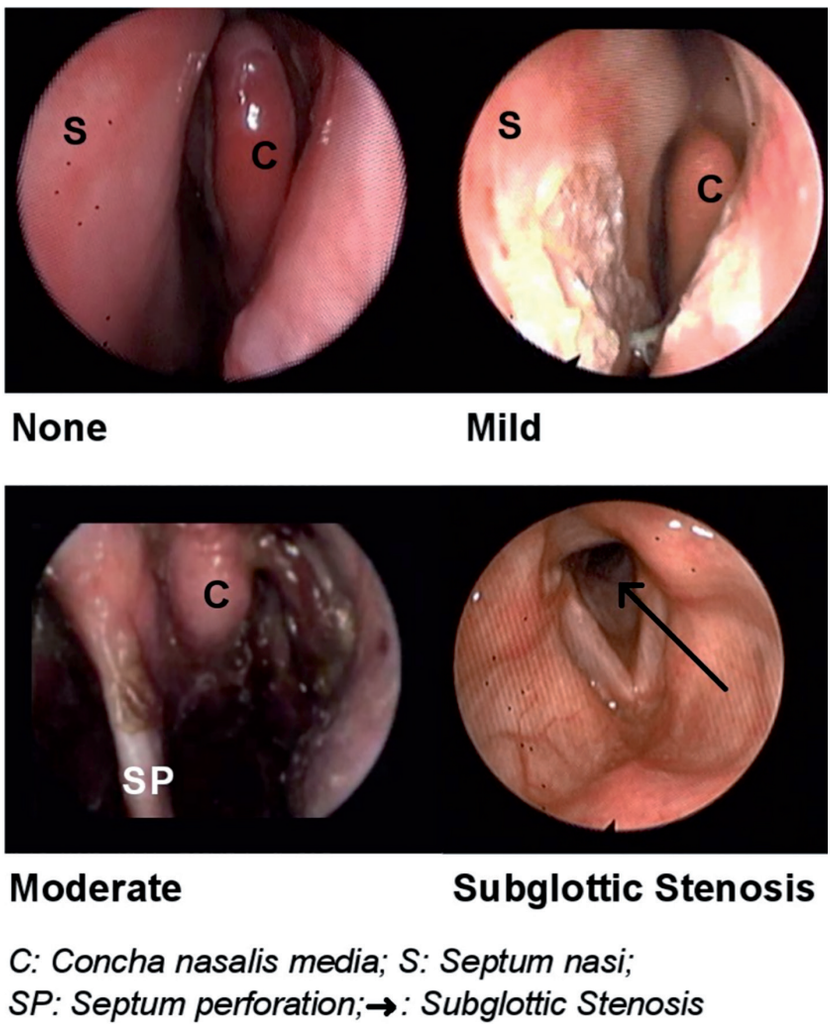

Fig. 2. Endoscopic findings.

by using the classification of the European Vasculitis Study Group (EUVAS) (24). GPA activity stages (remission; response; major-/minor- relapse; refractory disease; low-activity disease stage) were defined by using the classification of the EULAR (5). Systemic disease activity was measured using the BVASv.3 (25). Systemic disease damage was measured using the VDI (9). Systemic inflammation was estimated by the serum levels of c-reactive protein, erythrocyte sedimentation rate and white blood cell count. 43 of 47 patients received glucocorticoid therapy with prednisolone (mean daily dosage 6.8 mg, max 65 mg, min 2 mg, SD=10.4 mg). 42 patients additionally received immunosuppressive therapy in different combinations and dosages: methotrexate (n=16), azathioprine (n=13), leflunomide (n=6), cotrimoxazole (n=3), mycophenolate mofetil (n=2), rituximab (n=7).

Reliability and multi-method agreement trial

The trial took place in a monocentric setting in the Department of Otorhinolaryngology, Head and Neck Surgery of the Kiel University, Germany. GPA activity and damage in the ENT tract were evaluated by a high-experienced single reference assessor (>200 GPA patients seen before) and three otorhinolaryngologists (>50 GPA patients seen before). GPA/ENT general activity was measured in dichotomy (yes/no) and in grading (none/mild/moderate/high). Damage was measured in dichotomy. Specific activity and damage evaluation of the ear, the nose and the throat was measured in dichotomy.

A single assessor reference (R0) was created: 47 endoscopic videos of the nasal cavity, pharynx, larynx, ears and oral cavity were given in random order. In a forced multiple choice mode, the reference assessor was asked to appraise ENT/GPA activity (dichotomy and grading) and damage (dichotomy) on the basis of the endoscopic videos. After adding further ENTAS 2 data (anamnesis and the results of ENT diagnostics) of the same 47 patients, activity and damage were re-evaluated by R0. The re-evaluations equate ENTAS 2 evaluations.

To prove the intra- and inter-rater reliability of the ENTAS 2, the appraisal was repeated by three physicians (R1-R3) who also evaluated activity and damage firstly due to video data, and secondly after adding anamnestic and diagnostic data. The assessment was rerun by R1-R3 in randomly changed patient order after 6 weeks (T1=07/2015; T2=09/2015). To verify the impact of endoscopy within the ENTAS 2, the multi-method agreement between endoscopic video-data based evaluations and added-data re-evaluations (≙ ENTAS 2 evaluations) was determined for the reference (R0) and the physicians (R1-R3).

Statistic analysis

Statistic analysis was performed by using SPSS v. 22.0 (IBM Corp.) and especially for the intra- and inter-rater-reliability analysis by using the package irr within the R program (R Core Team). To describe the patient population, mean values (M), ± standard deviation (SD), maximum and minimum were used. Percent agreement and Kappa (κ) statistics were used to describe reliability and multi-method agreement. Due to rating scale variations (nominal/ordinal), different Kappas were used: Dichotomous intra-rater reliability/multi-method agreement: Cohen's Kappa (26); dichotomous inter-rater reliability: Fleiss' Kappa (27); reliability/multi-method agreement in grading: Weighted Kappa (28).

Results

Patient characteristics

Mean time from diagnosis until study entry was M=6 years (range 0–21;

SD=6). Mean time from first manifestations of GPA until study entry was M=8 years (range 0–34; SD=8). C-ANCA was positive in 42 patients (89%). Forty (85%) patients were PR3-ANCA positive. GPA was biopsy proven in 43 patients (91%). In 14 patients, the disease stage was classified as localised GPA, in 30 patients as early systemic and in two patients as generalised. Furthermore, 27 patients were classified in remission, 9 in response, 7 showed a minor relapse, one was refractory and two had low-activity. Mean BVASv.3 was M=2.3 (range 0–14; SD=3.8). Mean VDI was M=2.5 (range 0–7; SD=1.5). Patient characteristics are shown in Table I.

7/47 patients (14.9%) of the cohort showed a subglottic stenosis (SGS). Video-data based SGS detection (present/absent) was marked as "undeclared" in 4.6% of the estimations. Additionally, SGS was evaluated in typing (circumferential/ spiral/ web) and grading (<50%/ 51–70%/ 71–99%/ 100%). In SGS typing and grading evaluation, "undeclared" was marked in 10.6%.

Table I. Patients' characteristics.

	GPA (n=47)
Mean age in years (range; SD)	56 (20-85; SD=15)
Sex	Female=30; Male=17
Biopsy proof of GPA	43 (91%)
c-ANCA positive	42 (89%)
PR3-ANCA positive	40 (85%)
Mean CRP in mg/dl (range; SD)	0.35 (0-7.7; 14) Standard value:<1mg/dl
Mean ESR in mm after the first hour (range, SD)	16 (2-70; 20.5)
Mean WBC per nl (range; SD)	8 (4.8-23.3; 3.4) Standard value:4-10/nl
BVASv.3 (range; SD)	2.3 (0-14; 3.8)
VDI (range; SD)	2.5 (0-7; 1.5)
EUVAS Disease stage (mD=1)	
Localised	14 (30%)
Early Systemic	30 (64%)
Generalised	2 (4%)

GPA: granulomatosis with polyangiitis; SD: standard deviation; CRP: C-reactive protein; ESR: erythrocyte sedimentation rate; WBC: white blood cell count; BVASv.3: Birmingham Vasculitis Activity Score version 3; VDI: Vasculitis Damage Index; EUVAS: European Vasculitits Study Group; mD: missing data.

Table II. ENTAS 2 intra-rater reliability T1 *vs*. T2.

Rating category	Activity - Intra-rater reliability		Damage - Intra-rater reliability	
	Ø R1-R3 (%)	Ø R1-R3 (κ)	Ø R1-R3 (%)	Ø R1-R3 (κ)
General (yes/no)	80.7	*0.56*	87.8	*0.74*
Grading	72.8	*0.41*		
Ear	90	*0.68*	91.3	-
Nose	86.4	*0.66*	88.9	*0.78*
Throat	96.4	-	92.9	-

R1-R3: physicians 1–3; ø: average value; %: Percent agreement; κ: Cohen's Kappa; - : Cohen's Kappa negative or 0.

Reliability

Note: Percent agreement alone should not be used as the only way to analyse reliability (29, 30). Also, Kappa alone has little value due to its dependency of the balance of marginal totals (31, 32). Therefore, the interpretation of reliability results as a combination of both percent agreement and kappa statistics is recommended (33). Negative Kappa is possible, but should not be interpreted (34).

Intra-rater reliability T1 to T2

ENTAS 2 activity evaluations at T1 and T2 had 80.7% (κ=0.56) intra-rater reliability in dichotomy (yes/no) and 72.8% (κ=0.41) in grading. ENTAS 2 damage evaluation (yes/no) intra-rater reliability between T1 and T2 was 87.8% (κ=0.74). Intra-rater reliability results are shown in Table II.

Inter-rater reliability between reference and physicians

In ENTAS 2 activity evaluation, dichotomous inter-rater reliability was 84.8% (κ=0.61) for R0 *vs*. R1, 84.8% (κ=0.66) for R0 *vs*. R2 and 76.6% (κ=0.51) for R0 *vs*. R3. Activity grading inter-rater reliability for R0 *vs*. R1 was 78.3% (κ=0.48), for R0 *vs*. R2 71.74% (κ=0.42) and for R0 *vs*. R3 61.7% (κ=0.29). In damage evaluation, inter-rater reliability was 80.4% (κ=0.62) for R0 *vs*. R1, 80.4% (κ=0.62) for R0 *vs*. R2 and 80.9% (κ=0.63) for R0 *vs*. R3.

Inter-rater reliability between the physicians

ENTAS 2 activity evaluation inter-rater reliability (R1-R3) at T1 was 62.2% (κ=0.43) in dichotomy and 51.1% (κ=0.29) in grading. At T2, it was 68.2% (κ=0.48) in dichotomy and 55.32% (κ=0.33) in grading. Inter-rater reliability (R1-R3) of ENTAS 2 damage evaluation was 84.4% (κ=0.79) at T1 and 72.5% (κ=0.64) at T2. Inter-rater reliability results are shown in Table III.

Specific activity and damage evaluations (ear/nose/throat) of the physicians

Intra-rater reliability of specific activity evaluations of the ear was 90% (κ=0.68), of the nose 86.4% (κ=0.66) and of the throat 96.4% (κ≤0). Intra-rater reliability of specific damage evaluations of the ear was 91.3% (κ≤0), of the nose 88.9% (κ 0.78) and of the throat 92.9% (κ≤0). Inter-rater reliability of specific activity evaluations of the ear was 74.47–84.44 (κ=0.42–0.68), of the nose 66.67–78.72% (κ=0.47–0.57) and of the throat 91.49–93.33% (κ≤0). Inter-rater reliability of specific damage evaluations of the ear was 87.23–91.11 (κ=0.47–0.57), of the nose 84.44–89.36% (κ=0.78–0.84) and of the throat 74.47-88.89% (κ≤0). Intra- and inter-rater reliability of specific activity and damage evaluations are shown in Table II and Table III.

Multi-method agreement

To prove the impact of endoscopy based GPA activity and damage evaluations within the ENTAS 2, the multi-method agreement between endoscopic

Table III. ENTAS 2 inter-rater reliability.

Rating category Activity - Inter-rater reliability									
		R1-R3 (%)	R1-R3 (κ)	R0/R1 (%)	R0/R1 (κ)	R0/R2 (%)	R0/R2 (κ)	R0/R3 (%)	R0/ R3 (κ)
General (yes/no)	T1	62.22	*0.43*	84.78	*0.61*	84.78	*0.66*	76.60	*0.51*
	T2	68.22	*0.48*						
Grading	T1	51.11	*0.29*	78.26	*0.48*	71.74	*0.42*	61.70	*0.29*
	T2	55.32	*0.33*						
Ear	T1	84.44	*0.61*	93.48	*0.54*	89.13	*0.40*	85.11	*0.32*
	T2	74.47	*0.42*						
Nose	T1	66.67	*0.47*	84.78	*0.58*	91.30	*0.80*	82.98	*0.62*
	T2	78.72	*0.57*						
Throat	T1	93.33	-	95.65	-	95.65	-	95.74	-
	T2	91.49	-						
Rating category Damage - Inter-rater reliability									
		R1-R3 (%)	R1-R3 (κ)	R0/R1 (%)	R0/R1 (κ)	R0/R2 (%)	R0/R2 (κ)	R0/R3 (%)	R0/ R3 (κ)
General (yes/no)	T1	84.44	*0.79*	80.43	*0.62*	80.43	*0.62*	80.85	*0.63*
	T2	74.47	*0.64*						
Ear	T1	91.11	*0.47*	93.48	*0.37*	93.48	*0.37*	95.74	*0.48*
	T2	87.23	*0.52*						
Nose	T1	84.44	*0.78*	93.33	*0.86*	93.33	*0.86*	93.48	*0.86*
	T2	89.36	*0.84*						
Throat	T1	88.89	-	73.91	*0.11*	67.39	-	70.21	*0.1*
	T2	74.47	-						

R0: reference; R1-R3: physicians 1-3; ø: average value; T1: first evaluation; T2: second evaluation; %: Percent agreement; κ: Fleiss' Kappa; - : Fleiss' Kappa negative or 0

Table IV. Multi-method agreement video-data based evaluation *versus* ENTAS 2.

Rating category	Activity - Multi-method agreement				Damage - Multi-method agreement			
	R0 (%)	R0 (ê)	ø R1-R3 (%)	ø R1-R3 (ê)	R0 (%)	R0 (ê)	ø R1-R3 (%)	ø R1-R3 (ê)
General (yes/no)	91.5	*0.79*	94.7	*0.87*	76.6	*0.54*	98.2	*0.96*
Grading	91.5	*0.79*	94.3	*0.85*				
Ear	100	*1.00*	94.3	*0.79*	97.9	*0.60*	99.6	*0.97*
Nose	93.6	*0.84*	97.9	*0.94*	93.5	*0.86*	97.9	*0.95*
Throat	97.9	-	99.6	-	74.5	*0.30*	99.7	-

R0: reference; R1-R3: physicians 1–3; %: Percent agreement ê: Cohen's Kappa; - : Cohen's Kappa negative or 0; ø: average value.

video-data based evaluations and re-evaluations after adding anamnestic and diagnostic data was determined (re-evaluation ≙ ENTAS 2 evaluation).

R0 multi-method agreement

R0 multi-method agreement between video-data based GPA/ENT activity evaluations and ENTAS 2 activity evaluations was 91.5% (κ=0.79) in dichotomy and 91.5% (κ=0.78) in grading. R0 general multi-method agreement of damage evaluations was 76.6% (κ=0.54). R0 multi-method agreement of specific activity and damage evaluations of the ear and the nose was >90% (κ=0.6–1.0). R0 multi-method agreement was 97.9% (κ≤ 0) in throat activity and 74.5% (κ=0.3) in throat damage evaluations.

R1-R3 multi-method agreement

R1-R3 multi-method agreement of activity evaluations was 94.7% (κ=0.87) in dichotomy and 94.3% (κ=0.85) in grading. Mean R1-R3 multi-method agreement was 98.2% (κ=0.96) in damage evaluations. Mean R1-R3 multi-method agreement of specific activity and damage evaluations of the ear and the nose was >90% (κ=0.84–0.95). Mean R1-R3 multi-method agreement was 99.6% (κ≤0) in throat activity and 99.7% (κ≤0) in throat damage evaluations. Multi-method agreement results are shown in Table IV.

Discussion

There is no objective marker or item that correlates well with activity and damage of GPA in the ENT region described yet (35). The 2016 EULAR/ERA-EDTA recommendations state the use of structural clinical assessment to inform decisions on changes in treatment of ANCA-associated vasculitis (3). It is necessary to validate activity and damage by the use of score-systems, which are obviously dependent on observer evaluations. Hence, the repeatability (intra-rater reliability) and reproducibility (inter-rater reliability) of these evaluations are of particular interest. ENT assessment is essential to evaluate GPA activity and damage, as the ENT region shows some of the most frequent GPA manifestations (36). Trials to create structured clinical ENT assessment schemes have been described (21, 37), but the intra- and inter-rater reliability of a full score-system based ENT assessment have not been determined yet.

The present study describes the modified ENT assessment score based on anamnesis, endoscopic and diagnostic data (ENTAS 2). It showed high intra-rater reliability in dichotomous (yes/no) activity (80.7%, κ=0.56) and even in activity grading (none/mild/moderate/high) (72.8%, κ=0.41). For the BVASv.3 ENT sub-score, an intra-rater reliability of κ=0.94 was described (25). However, there is no data published about the level of experience of the ENT sub-score raters, which may have affected the reliability results. The ENTAS 2 showed high intra-rater reliability in damage evaluation (87.8%, κ=0.74). This is similar to the findings

of Suppiah *et al*., who described an intra-rater reliability of κ=0.79 for the VDI/ENT sub-score and κ=0.78 for the Combined Damage Assessment Index (CDA) ENT part (38). Thus, there is probably not enough data about the patient characteristics of the cohort included in that intra-rater reliability analysis given. The intra-rater reliability was only determined by including 14 patients for the VDI/ENT part and 15 patients for the CDA/ENT part. The dimension of homogeneity of that small cohort may have influenced the results.
R1-R3 showed high percentages of agreement towards R0 both in dichotomous activity and damage evaluation (76.6–84.8%, κ=0.51–0.63), whereas the activity grading evaluations of R1-R3 showed a greater variation in agreement towards the reference (61.7–78.3%, κ=0.29–0.43). Inter-rater reliability between R1-R3 showed good agreement in dichotomous activity evaluation (T1: 62.3%, κ=0.43; T2: 68.2%, κ=0.48). Unfortunately, activity grading evaluation showed little reliable inter-rater agreement (T1: 51.11%, κ=0.29, T2: 55.32%, κ=0.33). The BVASv.3 ENT sub-score showed an inter-rater reliability of κ=0.89 (25). Though, besides the previously mentioned lack of rater information, there is neither data about the disease stages nor about the activity stages of the patient cohort given. The distribution of disease and activity stages also may have affected the reliability result. In damage evaluation, the ENTAS 2 led to a high percentage of inter-rater agreement (T1: 84.4%, κ=0.79; T2: 74.5%, κ=0.64). The VDI showed lower inter-rater reliability (κ=0.41), and the CDA ENT part also showed a lower inter-rater kappa value (κ=0.59) (9, 38). The Disease Activity Index for ENT involvement showed good sensitivity and specifity, but no intra- and inter-rater reliability has been determined yet (18).
In GPA/ENT assessment, endoscopy has a central role. By the use of it, not only mucosa can be analysed, but also damages such as subglottic stenoses can be detected.
As nasal mucosa affection is present in GPA in up to 60–90% (10), rhinoscopy is probably one of the most important assessments in GPA diagnostics. Although rhinoscopy is a routine ENT examination, it is challenging to standardise. Lack in inter-rater reliability was noted in different rhinoscopy studies. In a study about chronic rhinosinusitis, values of κ=0.42 in the evaluation of "nasal discharge" and values of κ=0.24-0.28 in the evaluation of "nasal obstruction" were reported (39). In a Turkish study, "turbinate colour" evaluation in allergic rhinitis showed little inter-rater agreement (κ=0.38) (40). On the other hand, Annamalai *et al*. reported high agreement for "crusting" (κ=0.62) and "nasal discharge" (κ=0.84) and recommended the use of a standardised endonasal score-system (41). The ENTAS 2 is a useful score-system to evaluate nasal mucosa activity in GPA patients. In the study, the inter-rater reliability considering ENTAS 2 nose activity evaluations showed high agreement (66.67–78.72%, κ=0.47-0.57). This is similar to the inter-rater reliability findings of Garske *et al*. who estimated reliability of the first version ENTAS endonasal GPA activity evaluation by using images made by rhinoscopy (κ=0.50–0.62) (20).
High specifity of selected ENT items and precise item descriptions may result in improvement of intra- and inter-rater reliability of activity evaluation in scoring systems. Nose items of both BVASv.3 and VDI are probably not specific enough. For example, in the BVASv.3, one item of nasal activity is described as "Light or dark brown crusts frequently obstructing the nose" (25). In the VDI, nasal damages are described as "Nasal blockage or chronic discharge or crusting" (9). However, endonasal crusting has not been verified to be a specific sign of activity as its presence correlates both with disease activity and infection (18, 42). Additionally, crusting has been described in wide nasal cavities and dry noses induced by GPA (43, 44).
This study showed a high percentage of agreement in inter-rater reliability of ENTAS 2 throat activity evaluation (91.49–93.33%). Unfortunately, in throat damage evaluation was greater variety in inter-rater reliability (74.5-88.9%). This may correlate with challenges in the interpretation of larynx findings. Subglottic stenosis is the most common throat damage in GPA and is featured in the VDI (9,36). However, detection of SGS by endoscopy is demanding as the view on the trachea can be insufficient (45, 46). In the present study, 4.6% of endoscopy based SGS evaluations (absent/present) were marked as "undeclared". Endoscopy based SGS typing (circumferential/ spiral/ web) and grading (<50%/51–70%/71–99%/100%) evaluation was tagged as "undeclared" in 10.6% of the cases. Brook *et al*. stated a wide variation in agreement of interpretation of laryngoscopy findings and a dependency on observer experience (47). Therefore, complementing tools to diagnose SGS may be useful to standardise GPA/ENT damage evaluations. Solans-Laque *et al*. recommended the use of 3D-CT of the trachea and virtual bronchoscopy (48). Klink *et al*. described magnetic resonance imaging (MRI) as being a "promising tool" to detect and grade SGS in GPA (49). Flexible 'chip-on-the-tip' laryngoscopy can lead to a higher image quality than conventional flexible endoscopy (50). However, the technique has not been analysed in the use of mucosa estimation yet.
To verify the impact of endoscopy within a GPA/ENT score-system, it is relevant to determine the agreement between endoscopy and complete score-system based activity and damage evaluations. Multi-method agreement between video endoscopy based evaluations and ENTAS 2 based evaluations was high in the present study. Accordingly, in most cases GPA activity and damage were not evaluated differently whether anamnestic and diagnostic data were added or only endoscopy was presented. This may emphasise the decisive impact of endoscopy within the ENTAS 2 evaluations.
Sole exception of the multi-method agreement results was a greater variability in throat damage evaluations. Whereas the reference assessor video based *versus* added-data throat damage evaluations had an agreement of 74.5%, the mean agreement of the physicians' evaluation was 99.7%. There were differences in handling with anamnestic

data. Observers reacted differently on the addition of the anamnestic data "stridor", "voice changes" and "dyspnea". In some cases, observers changed their evaluation of throat damage from no to yes in the presence of these data and vice-versa in the abstinence of it. However, in other cases with a similar anamnestic profile, no throat evaluation changes were done, which may have pointed again to the challenge of diagnosis and evaluation of GPA throat involvement. The findings might correlate with observer uncertainty in the previous endoscopic throat evaluation. Sufficient visualisation of the trachea is essential in patients with stridor/breathing difficulties to detect GPA damages such as the SGS (45).

Limitations

There may be limitations of the reliability results of this study. A single assessor estimation was used as reference. Although the assessor was high-experienced in evaluation of GPA, the results remain subjective and observer-dependent. The video endoscopy evaluations were dependent on the video quality. All patients received immunosuppressant therapy, the majority was in remission or response and no major relapse was present. The results may have been different in a more heterogeneous observational cohort, as important variability between patients with ANCA-associated vasculitis enrolled in clinical trials and in observational cohorts has been described (51). Order effects such as the practice or the fatigue effect may have influenced the reliability results (52). Additionally, the results may have been more rater-independent with a higher number of physicians.

Conclusion

The ENTAS 2 showed high intra-rater reliability in activity and damage evaluation. In inter-rater reliability, the ENTAS 2 showed high agreement in dichotomous activity and damage evaluation, but low inter-rater reliability in activity grading evaluation. There was high multi-method agreement between video endoscopy and ENTAS 2 evaluations, except for throat damage. Our data lead to the conclusion that the ENTAS 2 is a reliable instrument in the evaluation of GPA/ENT activity and damage.

Acknowledgments

We thank M. Gonzalez, H. Brauer and J. Gauerke for their participation in the appraisals. We thank J. Hedderich for his statistical advice. We thank F. Moosig and J. Schirmer for their help in the systemic patient characterisations. We deeply acknowledge the participation of every patient.

References

1. WOJCIECHOWSKA J, KRAJEWSKI W, KRAJEWSKI P, KRECICKI T: Granulomatosis with polyangiitis in otolaryngologist practice: a review of current knowledge. *Clin Exp Otorhinolaryngol* 2016; 9: 8-13.
2. ELEFANTE E, TRIPOLI A, FERRO F, BALDINI C: One year in review: systemic vasculitis. *Clin Exp Rheumatol* 2016; 34 (Suppl. 97): S1-6.
3. YATES M, WATTS RA, BAJEMA IM *et al*.: EULAR/ERA-EDTA recommendations for the management of ANCA-associated vasculitis. *Ann Rheum Dis* 2016; 75: 1583-94.
4. SUPPIAH R, MUKHTYAR C, FLOSSMANN O *et al*.: A cross-sectional study of the Birmingham Vasculitis Activity Score version 3 in systemic vasculitis. *Rheumatology* (Oxford) 2011; 50: 899-905.
5. HELLMICH B, FLOSSMANN O, GROSS WL *et al*.: EULAR recommendations for conducting clinical studies and/or clinical trials in systemic vasculitis: focus on anti-neutrophil cytoplasm antibody-associated vasculitis. *Ann Rheum Dis* 2007; 66: 605-17.
6. LUQMANI RA, BACON PA, MOOTS RJ *et al*.: Birmingham Vasculitis Activity Score (BVAS) in systemic necrotizing vasculitis. *QJM* 1994; 87: 671-8.
7. WHITING-O'KEEFE QE, STONE JH, HELLMANN DB: Validity of a vasculitis activity index for systemic necrotizing vasculitis. *Arthritis Rheum* 1999; 42: 2365-71.
8. KALLENBERG CG, TERVAERT JW, STEGEMAN CA: Criteria for disease activity in Wegener's granulomatosis: a requirement for longitudinal clinical studies. *APMIS* Suppl 1990; 19: 37-9.
9. EXLEY AR, BACON PA, LUQMANI RA *et al*.: Development and initial validation of the Vasculitis Damage Index for the standardized clinical assessment of damage in the systemic vasculitides. *Arthritis Rheum* 1997; 40: 371-80.
10. SROUJI IA, ANDREWS P, EDWARDS C, LUND VJ: Patterns of presentation and diagnosis of patients with Wegener's granulomatosis: ENT aspects. *J Laryngol Otol* 2007; 121: 653-8.
11. QUETZ J: Sattelnase bei M. Wegener: Rekonstruktive Rhinoplastik mit autologen Rippenknorpetransplantaten durch geschlossene Technik [Internet]. Düsseldorf: German Medical Science GMS Publishing House. 2008. Available from: http://www.egms.de/de/meetings/hnod2008/08hnod540.shtml. Access date: 12/28/2016
12. SEPEHR A, ALEXANDER AJ, CHAUHAN N, GANTOUS A: Detailed analysis of graft techniques for nasal reconstruction following Wegener granulomatosis. *J Otolaryngol Head Neck Surg* 2011; 40: 473-80.
13. GLUTH MB, SHINNERS PA, KASPERBAUER JL: Subglottic stenosis associated with Wegener's granulomatosis. *Laryngoscope* 2003; 113: 1304-7.
14. MARTINEZ DEL PERO M, JAYNE D, CHAUDHRY A, SIVASOTHY P, JANI P: Long-term outcome of airway stenosis in granulomatosis with polyangiitis (Wegener granulomatosis): an observational study. *JAMA Otolaryngol Head Neck Surg* 2014; 140: 1038-44.
15. WIERZBICKA M, TOKARSKI M, PUSZCZEWICZ M, SZYFTER W: The efficacy of submucosal corticosteroid injection and dilatation in subglottic stenosis of different aetiology. *J Laryngol Otol* 2016; 130: 674-9.
16. WOLTER NE, OOI EH, WITTERICK IJ: Intralesional corticosteroid injection and dilatation provides effective management of subglottic stenosis in Wegener's granulomatosis. *Laryngoscope* 2010; 120: 2452-5.
17. HOFFMAN GS, THOMAS-GOLBANOV CK, CHAN J, AKST LM, ELIACHAR I: Treatment of subglottic stenosis, due to Wegener's granulomatosis, with intralesional corticosteroids and dilation. *J Rheumatol* 2003; 30: 1017-21.
18. DEL PERO MM, CHAUDHRY A, RASMUSSEN N, JANI P, JAYNE D: A disease activity score for ENT involvement in granulomatosis with polyangiitis (Wegener's). *Laryngoscope* 2013; 123: 622-8.
19. FELICETTI M, CAZZADOR D, FACCIOLI C *et al*.: Clinical application of two different disease activity scores for ENT involvement in granulomatosis with polyangiitis. *Ann Rheum Dis* 2016; 75 (Suppl. 2): 1093.
20. GARSKE U, HAACK A, BELTRÁN O *et al*.: Intra- and inter-rater reliability of endonasal activity estimation in granulomatosis with polyangiitis (Wegener's). *Clin Exp Rheumatol* 2012; 30 (Suppl. 70): S22-8.
21. MARTINEZ DEL PERO M, RASMUSSEN N, CHAUDHRY A, JANI P, JAYNE D: Structured clinical assessment of the ear, nose and throat in patients with granulomatosis with polyangiitis (Wegener's). *Eur Arch Otorhinolaryngol* 2013; 270: 345-54.
22. LEAVITT RY, FAUCI AS, BLOCH DA *et al*.: The American College of Rheumatology 1990 criteria for the classification of Wegener's granulomatosis. *Arthritis Rheum* 1990; 33: 1101-7.
23. JENNETTE JC, FALK RJ, BACON PA *et al*.: 2012 revised International Chapel Hill Consensus Conference Nomenclature of Vasculitides. *Arthritis Rheum* 2013; 65: 1-11.
24. MOOSIG F, HOLLE JU, GROSS WL: Autoimmunvaskulitiden: Standards und Leitlinien nach EULAR und EUVAS [Autoimmune vasculitides. Standards and guidelines of EULAR and EUVAS]. *Internist* (Berl) 2009; 50: 298.
25. MUKHTYAR C, LEE R, BROWN D *et al*.: Modification and validation of the Birming-

ham Vasculitis Activity Score (version 3). *Ann Rheum Dis* 2009; 68: 1827-32.
26. COHEN J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 2037.
27. FLEISS JL: The measurement of interrater agreement. Statistical methods for rates and proportions. 2nd Ed., New York, John Wiley & Sons, 1981; 212-36.
28. SPITZER R, FLEISS JL, COHEN J: Mental Status Schedule - Properties of Factor-Analytically Derived Scales. *Arch Gen Psychatry* 1967; 16: 479-93.
29. LOMBARD M, SNYDER-DUCH JB: Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research* 2002; 28: 587.
30. JOYCE M: Picking the best intercoder reliability statistic for your digital activism content analysis [Internet]. 2013. Available from: http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis/. Access date: 12/28/2016
31. LANTZ CA, NEBENZAHL E: Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *J Clin Epidemiol* 1996; 49: 431-4.
32. FEINSTEIN AR, CICCHETTI DV: High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; 43: 543.
33. STEINIJANS VW, DILETTI E, BOMCHES B, GREIS C, SOLLEDER P: Interobserver agreement: Cohen's kappa coefficient does not necessarily reflect the percentage of patients with congruent classifications. *Int J Clin Pharmacol Ther* 1997; 35: 93-5.
34. SIM J, WRIGHT CC: The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Phys Ther 2005; 85: 257-68.
35. LUQMANI RA: Disease assessment in systemic vasculitis. *Nephrol Dial Transplant* 2015; 30 (Suppl. 1): i76-82.
36. LAUDIEN M, AMBROSCH P, TILL A, PODSCHUN R, LAMPRECHT P: Diagnosis, therapy and current research aspects of selected chronic inflammatory diseases with head and neck involvement [Diagnostik, Therapie und aktuelle Forschungsaspekte ausgewahlter chronisch-entzundlicher Erkrankungen mit Beteiligung im Kopf-Hals-Bereich]. *Z Rheumatol* 2008; 67: 397-406.
37. MORALES-ANGULO C, GARCIA-ZORNOZA R, OBESO-AGUERA S, CALVO-ALEN J, GONZALEZ-GAY MA: Ear, Nose and Throat Manifestations of Wegener's Granulomatosis (Granulomatosis with Polyangiitis) [Manifestaciones otorrinolaringologicas en pacientes con granulomatosis de Wegener (granulomatosis con poliangeitis)]. *Acta Otorrinolaringol Esp* 2012; 63: 206-11.
38. SUPPIAH R, FLOSSMAN O, MUKHTYAR C *et al*.: Measurement of damage in systemic vasculitis: a comparison of the Vasculitis Damage Index with the Combined Damage Assessment Index. *Ann Rheum Dis* 2011; 70: 80-5.
39. MCCOUL ED, SMITH TL, MACE JC *et al*.: Interrater agreement of nasal endoscopy in patients with a prior history of endoscopic sinus surgery. *Int Forum Allergy Rhinol* 2012; 2 453.
40. EREN E, AKTAS A, ARSLANOGLU S *et al*.: Diagnosis of allergic rhinitis: inter-rater reliability and predictive value of nasal endoscopic examination: a prospective observational study. *Clin Otolaryngol* 2013; 38: 481-6.
41. ANNAMALAI S, DAVIS J, KUBBA H: How subjective is nasal endoscopy? A study of interrater agreement using the Lund and Mackay scoring system. *Am J Rhinol* 2004; 18: 301-3.
42. LAUDIEN M: Nasal barrier dysfunction in Wegener's granulomatosis. *Clin Exp Rheumatol* 2010; 28 (Suppl. 57): S3-4.
43. HILDENBRAND T, WEBER RK, BREHMER D: Rhinitis sicca, dry nose and atrophic rhinitis: a review of the literature. *Eur Arch Otorhinolaryngol* 2011; 268: 17-26.
44. HUIZING EH, DE GROOT JAM: Functional Reconstructive Nasal Surgery. Stuttgart, New York, Thieme, 2003: 285-6.
45. RASMUSSEN N: L24. Local treatments of subglottic and tracheal stenoses in granulomatosis with polyangiitis (Wegener's). *Presse Med* 2013; 42: 571-4.
46. SEAM N, FINKELSTEIN SE, GONZALES DA, SCHRUMP DS, GLADWIN MT: The workup of stridor: virtual bronchoscopy as a complementary technique in the diagnosis of subglottic stenosis. *Respir Care* 2007; 52: 337-9.
47. BROOK CD, PLATT MP, RUSSELL K, GRILLONE GA, ALIPHAS A, NOORDZIJ JP: Time to competency, reliability of flexible transnasal laryngoscopy by training level: a pilot study. *Otolaryngol Head Neck Surg* 2015; 152: 843-50.
48. SOLANS-LAQUE R, BOSCH-GIL J, CANELA M, LORENTE J, PALLISA E, VILARDELL-TARRES M: Clinical features and therapeutic management of subglottic stenosis in patients with Wegener's granulomatosis. *Lupus* 2008; 17: 832-6.
49. KLINK T, HOLLE J, LAUDIEN M *et al*.: Magnetic resonance imaging in patients with granulomatosis with polyangiitis (Wegener's) and subglottic stenosis. *MAGMA* 2013; 26: 281-90.
50. SCHROCK A, STUHRMANN N, SCHADE G: Flexible 'chip-on-the-tip' endoscopy for larynx diagnostics [Flexible Chip-on-the-tip-Endoskopie zur Larynxdiagnostik]. *HNO* 2008; 56: 1239-42.
51. PAGNOUX C, CARETTE S, KHALIDI NA *et al*.: Comparability of patients with ANCA-associated vasculitis enrolled in clinical trials or in observational cohorts. *Clin Exp Rheumatol* 2015; 33 (Suppl. 89): S77-83.
52. COZBY PC: Methods in behavioral research. McGraw-Hill Higher Education, 10th Edition, 2009; 155.